Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 25
March 31, 2023

Sampling methods

Metropolis Hastings algorithm[1]

[1] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Sampling methods

Metropolis Hastings algorithm[1]

Suppose the current state is $\mathbf{x}_t$ and the next step is $\mathbf{x}_{t+1}$ with probability $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t)$, $g(\mathbf{x})$ is the proposal distribution

[1]Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Sampling methods

Metropolis Hastings algorithm[1]

- Suppose the current state is $\mathbf{x}_t$ and the next step is $\mathbf{x}_{t+1}$ with probability $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t)$, $g(\mathbf{x})$ is the proposal distribution
- A commonly used $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1}|\,\mathbf{x}_t, \mathbf{\Sigma})$ i.e. a Gaussian distribution 'centered' on the current state, also called random walk Metropolis algorithm

[1] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Sampling methods

Metropolis Hastings algorithm[1]

- Suppose the current state is $\mathbf{x}_t$ and the next step is $\mathbf{x}_{t+1}$ with probability $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t)$, $g(\mathbf{x})$ is the proposal distribution
- A commonly used $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1}|\,\mathbf{x}_t, \mathbf{\Sigma})$ i.e. a Gaussian distribution 'centered' on the current state, also called random walk Metropolis algorithm
- If we choose the proposal independent of the old state i.e. if we set $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t) = g(\mathbf{x}_{t+1})$ then it is called independence sampler

[1] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Sampling methods

Metropolis Hastings algorithm[1]

- Suppose the current state is $\mathbf{x}_t$ and the next step is $\mathbf{x}_{t+1}$ with probability $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t)$, $g(\mathbf{x})$ is the proposal distribution
- A commonly used $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t+1}|\,\mathbf{x}_t, \boldsymbol{\Sigma})$ i.e. a Gaussian distribution 'centered' on the current state, also called random walk Metropolis algorithm
- If we choose the proposal independent of the old state i.e. if we set $g(\mathbf{x}_{t+1}|\,\mathbf{x}_t) = g(\mathbf{x}_{t+1})$ then it is called independence sampler

Question How to choose $\boldsymbol{\Sigma}$ and what is the acceptance probability?

[1] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. MIT press.

Sampling methods

Recall Our original problem was sampling from a probability distribution

$$p(\mathbf{x}) = \frac{f(\mathbf{x})}{NC},$$

where the normalizing constant NC is hard to compute, and the goal was to sample from a known suitable distribution which is easy to simulate.

Sampling methods

Recall Our original problem was sampling from a probability distribution

$$p(\mathbf{x}) = \frac{f(\mathbf{x})}{NC},$$

where the normalizing constant NC is hard to compute, and the goal was to sample from a known suitable distribution which is easy to simulate. If we manage to determine NC then $p(\mathbf{x}) = \pi$, the target distribution.

Sampling methods

Recall Our original problem was sampling from a probability distribution

$$p(\mathbf{x}) = \frac{f(\mathbf{x})}{NC},$$

where the normalizing constant NC is hard to compute, and the goal was to sample from a known suitable distribution which is easy to simulate. If we manage to determine NC then $p(\mathbf{x}) = \pi$, the target distribution.

Suppose NC is hard to compute (the method is obviously valid when we are able to compute NC)

Sampling methods

Since the target distribution is the stationary and the MC is reversible, the balance equation is satisfied i.e. for any two states $\pi_i p_{ij} = \pi_j p_{ji}$.

Sampling methods

Since the target distribution is the stationary and the MC is reversible, the balance equation is satisfied i.e. for any two states $\pi_i p_{ij} = \pi_j p_{ji}$. Thus we have

$$p(\mathbf{x}_i)p_{ij} = p(\mathbf{x}_j)p_{ji}$$

that is

$$\frac{f(\mathbf{x}_i)}{NC} g(\mathbf{x}_j | \mathbf{x}_i) a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{NC} g(\mathbf{x}_i | \mathbf{x}_j) a(\mathbf{x}_j \to \mathbf{x}_i)$$

where $a(s_i \to s_j)$ denotes the accept probability from s_i state to s_j.

Sampling methods

Since the target distribution is the stationary and the MC is reversible, the balance equation is satisfied i.e. for any two states $\pi_i p_{ij} = \pi_j p_{ji}$. Thus we have

$$p(\mathbf{x}_i)p_{ij} = p(\mathbf{x}_j)p_{ji}$$

that is

$$\frac{f(\mathbf{x}_i)}{NC} g(\mathbf{x}_j|\, \mathbf{x}_i) a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{NC} g(\mathbf{x}_i|\, \mathbf{x}_j) a(\mathbf{x}_j \to \mathbf{x}_i)$$

where $a(s_i \to s_j)$ denotes the accept probability from s_i state to s_j. Thus

$$\frac{a(\mathbf{x}_i \to \mathbf{x}_j)}{a(\mathbf{x}_j \to \mathbf{x}_i)} = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\, \mathbf{x}_j)}{g(\mathbf{x}_j|\, \mathbf{x}_i)}$$

Sampling methods

Since the target distribution is the stationary and the MC is reversible, the balance equation is satisfied i.e. for any two states $\pi_i p_{ij} = \pi_j p_{ji}$. Thus we have

$$p(\mathbf{x}_i)p_{ij} = p(\mathbf{x}_j)p_{ji}$$

that is

$$\frac{f(\mathbf{x}_i)}{NC} g(\mathbf{x}_j|\mathbf{x}_i) a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{NC} g(\mathbf{x}_i|\mathbf{x}_j) a(\mathbf{x}_j \to \mathbf{x}_i)$$

where $a(s_i \to s_j)$ denotes the accept probability from s_i state to s_j. Thus

$$\frac{a(\mathbf{x}_i \to \mathbf{x}_j)}{a(\mathbf{x}_j \to \mathbf{x}_i)} = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\mathbf{x}_j)}{g(\mathbf{x}_j|\mathbf{x}_i)}$$

Next task is to decide the values of the accept probabilities from this equation

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i | \mathbf{x}_j)}{g(\mathbf{x}_j | \mathbf{x}_i)} < 1$

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} < 1$

Case II. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} > 1$

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} < 1$

Case II. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} > 1$

For Case I. set $a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = 1$

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} < 1$

Case II. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} > 1$

For Case I. set $a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = 1$

For case II. set $a(\mathbf{x}_i \to \mathbf{x}_j) = 1$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_j)} \cdot \frac{g(\mathbf{x}_j|\,\mathbf{x}_i)}{g(\mathbf{x}_i|\,\mathbf{x}_j)}$

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} < 1$

Case II. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} > 1$

For Case I. set $a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = 1$

For case II. set $a(\mathbf{x}_i \to \mathbf{x}_j) = 1$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_j)} \cdot \frac{g(\mathbf{x}_j|\,\mathbf{x}_i)}{g(\mathbf{x}_i|\,\mathbf{x}_j)}$

Now assuming the proposal distribution to be symmetric i.e. $g(\mathbf{x}|\,\mathbf{y}) = g(\mathbf{y}|\mathbf{x})$, we set

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)}\right\}$$

Sampling methods

Two cases arise:

Case I. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} < 1$

Case II. $\frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)} > 1$

For Case I. set $a(\mathbf{x}_i \to \mathbf{x}_j) = \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = 1$

For case II. set $a(\mathbf{x}_i \to \mathbf{x}_j) = 1$ and $a(\mathbf{x}_j \to \mathbf{x}_i) = \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_j)} \cdot \frac{g(\mathbf{x}_j|\,\mathbf{x}_i)}{g(\mathbf{x}_i|\,\mathbf{x}_j)}$

Now assuming the proposal distribution to be symmetric i.e. $g(\mathbf{x}|\,\mathbf{y}) = g(\mathbf{y}|\mathbf{x})$, we set

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)}\right\}$$

This is due to Metropolis Algo

Sampling methods

If the proposal distribution is NOT symmetric (asymmetric) i.e. $g(\mathbf{x}|\,\mathbf{y}) \neq g(\mathbf{y}|\mathbf{x})$ then define

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}\right\}$$

Sampling methods

If the proposal distribution is NOT symmetric (asymmetric) i.e. $g(\mathbf{x}|\,\mathbf{y}) \neq g(\mathbf{y}|\mathbf{x})$ then define

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}\right\}$$

This is due to Hastings Algo

Question What about the convergence? What about the accuracy?

Sampling methods

If the proposal distribution is NOT symmetric (asymmetric) i.e. $g(\mathbf{x}|\,\mathbf{y}) \neq g(\mathbf{y}|\mathbf{x})$ then define

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}\right\}$$

This is due to Hastings Algo

Question What about the convergence? What about the accuracy?

Question Does it solve our original problem? What is next?

Sampling methods

If the proposal distribution is NOT symmetric (asymmetric) i.e. $g(\mathbf{x}|\,\mathbf{y}) \neq g(\mathbf{y}|\mathbf{x})$ then define

$$a(\mathbf{x}_i \to \mathbf{x}_j) = \min\left\{1, \frac{f(\mathbf{x}_j)}{f(\mathbf{x}_i)} \cdot \frac{g(\mathbf{x}_i|\,\mathbf{x}_j)}{g(\mathbf{x}_j|\,\mathbf{x}_i)}\right\}$$

This is due to Hastings Algo

Question What about the convergence? What about the accuracy?

Question Does it solve our original problem? What is next?

Gibbs Sampling

Sampling methods

Convergence and error bounds Since π is a stationary distribution, for any initial distribution μ and for any state $s_i \in S$, $P_\mu(X_t = s_i) \to \pi_i$ as $t \to \infty$.

In practice, the true distribution of X_t should closely approximate the target density π according to some measure

Sampling methods

Convergence and error bounds Since π is a stationary distribution, for any initial distribution μ and for any state $s_i \in S$, $P_\mu(X_t = s_i) \to \pi_i$ as $t \to \infty$.

- In practice, the true distribution of X_t should closely approximate the target density π according to some measure
- Then another question is: what parameter of the transition matrix controls the speed of the convergence

Sampling methods

Total variation metric: If f, g are absolutely continuous probability distributions on $\mathbb{R}^d$, $d \geq 1$ then

$$\rho(f, g) = \frac{1}{2} \int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x}$$

and for pmfs,

$$\rho(f, g) = \frac{1}{2} \sum_i |f(i) - g(i)|$$

Sampling methods

Total variation metric: If f, g are absolutely continuous probability distributions on $\mathbb{R}^d$, $d \geq 1$ then

$$\rho(f,g) = \frac{1}{2}\int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x}$$

and for pmfs,

$$\rho(f,g) = \frac{1}{2}\sum_i |f(i) - g(i)|$$

Separation distance: Let f, g be two pmfs. Then

$$D(f,g) = \sup_i \left(1 - \frac{f(i)}{g(i)}\right)$$

(the order matters)

Sampling methods

Total variation metric: If f, g are absolutely continuous probability distributions on $\mathbb{R}^d$, $d \geq 1$ then

$$\rho(f,g) = \frac{1}{2}\int |f(\mathbf{x}) - g(\mathbf{x})| d\mathbf{x}$$

and for pmfs,

$$\rho(f,g) = \frac{1}{2}\sum_i |f(i) - g(i)|$$

Separation distance: Let f, g be two pmfs. Then

$$D(f,g) = \sup_i \left(1 - \frac{f(i)}{g(i)}\right)$$

(the order matters)

Chi-square distance: suppose f, g have common support S, then

$$\chi^2(f,g) = \sqrt{\int_S \frac{(f-g)^2}{g}}$$

Sampling methods

Some more: Kolmogorov Metric, Hellinger Metric, Levy–Prokhorov Metric, Kullback–Leibler (KL) Distance, Wasserstein Distance, Bhattacharya Affinity, Rao's Geodesic Distances

Sampling methods

Some more: Kolmogorov Metric, Hellinger Metric, Levy–Prokhorov Metric, Kullback–Leibler (KL) Distance, Wasserstein Distance, Bhattacharya Affinity, Rao's Geodesic Distances

Question Are they metrics?

Sampling methods

Some more: Kolmogorov Metric, Hellinger Metric, Levy–Prokhorov Metric, Kullback–Leibler (KL) Distance, Wasserstein Distance, Bhattacharya Affinity, Rao's Geodesic Distances

Question Are they metrics?

Recall: Suppose π is the statonary distribution corresponding to a transition matrix P with the initial distribution μ. Then for the distribution of X_t for some fixed t, the ith entry of $\mu^{(t)}(=\mu P^t)$ is

$$P_\mu(X_t = i) = \sum_{j\in S} P(X_0 = j)[P^t]_{ji} = \sum_{j\in S} \mu_k p_{ji}^{(t)}$$

where $p_{ji}^{(t)} = [P^t]_{ji}$.

Sampling methods

The total variation distance between $\mu^{(t)}$ and the stationary distribution π with initial distribution μ is given by

$$\begin{aligned} \rho(\mu^{(t)}, \pi) &= \sup_{A \subseteq S} \left| P_{\mu^{(t)}}(A) - P_{\pi}(A) \right| \\ &= \sup_{A} \left| \sum_{i \in A} \sum_{j \in S} \mu_j p_{ji}^{(t)} - \sum_{i \in A} \pi_i \right| \end{aligned}$$

Sampling methods

The total variation distance between $\mu^{(t)}$ and the stationary distribution π with initial distribution μ is given by

$$\begin{aligned}\rho(\mu^{(t)}, \pi) &= \sup_{A \subseteq S} \left| P_{\mu^{(t)}}(A) - P_\pi(A) \right| \\ &= \sup_{A} \left| \sum_{i \in A} \sum_{j \in S} \mu_j p_{ji}^{(t)} - \sum_{i \in A} \pi_i \right|\end{aligned}$$

For real data simulation: we need to check if this measure is monotone decreasing wrt t

Sampling methods

The total variation distance between $\mu^{(t)}$ and the stationary distribution π with initial distribution μ is given by

$$\begin{aligned}\rho(\mu^{(t)}, \pi) &= \sup_{A \subseteq S} \left| P_{\mu^{(t)}}(A) - P_{\pi}(A) \right| \\ &= \sup_{A} \left| \sum_{i \in A} \sum_{j \in S} \mu_j p_{ji}^{(t)} - \sum_{i \in A} \pi_i \right|\end{aligned}$$

For real data simulation: we need to check if this measure is monotone decreasing wrt t

Separation distance:

$$D(\mu^{(t)}, \pi) = \sup_{i \in S} \left(1 - \frac{P_{\mu}(X_t = i)}{\pi_i} \right)$$

Sampling methods

Chi-square distance:

$$\chi^2(\mu^{(t)}, \pi) = \sum_{i \in S} \frac{(P_\mu(X_t = i) - \pi_i)^2}{\pi_i}$$

[2] Bremaud, P. (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Springer

Sampling methods

Chi-square distance:

$$\chi^2(\mu^{(t)}, \pi) = \sum_{i \in S} \frac{(P_\mu(X_t = i) - \pi_i)^2}{\pi_i}$$

Review of nonnegative matrix theory[2]

Perron-Frobenius Theorem Let A be a positive $r \times r$ matrix. Then there exists an eigenvalue $\lambda_1 > 0$ with algebraic and geometric multiplicity one such that $\lambda_1 > |\lambda_j|$ for any other eigenvalue λ_j. The eigenvector corresponding to the eigenvalue λ_1 is positive.

[2] Bremaud, P. (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Springer

Sampling methods

Chi-square distance:

$$\chi^2(\mu^{(t)}, \pi) = \sum_{i \in S} \frac{(P_\mu(X_t = i) - \pi_i)^2}{\pi_i}$$

Review of nonnegative matrix theory[2]

Perron-Frobenius Theorem Let A be a positive $r \times r$ matrix. Then there exists an eigenvalue $\lambda_1 > 0$ with algebraic and geometric multiplicity one such that $\lambda_1 > |\lambda_j|$ for any other eigenvalue λ_j. The eigenvector corresponding to the eigenvalue λ_1 is positive.

In particular, if A is stochastic then $\lambda_1 = 1$.

[2] Bremaud, P. (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Springer

Sampling methods

Chi-square distance:

$$\chi^2(\mu^{(t)}, \pi) = \sum_{i \in S} \frac{(P_\mu(X_t = i) - \pi_i)^2}{\pi_i}$$

Review of nonnegative matrix theory[2]

Perron-Frobenius Theorem Let A be a positive $r \times r$ matrix. Then there exists an eigenvalue $\lambda_1 > 0$ with algebraic and geometric multiplicity one such that $\lambda_1 > |\lambda_j|$ for any other eigenvalue λ_j. The eigenvector corresponding to the eigenvalue λ_1 is positive.

In particular, if A is stochastic then $\lambda_1 = 1$.

Question What is your conclusion about the transition matrix?

[2] Bremaud, P. (1999). Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Springer

Sampling methods

Notation Let the initial distribution μ be a one-point distribution for some $x \in S$ i.e. $p(X_0 = x) = 1$ i.e. $p(X_0 \in S \setminus \{x\}) = 0$. Then we denote

$$P^t(x, A) = P(X_t \in A | X_0 = x)$$

for any $A \subset S$

Sampling methods

Notation Let the initial distribution μ be a one-point distribution for some $x \in S$ i.e. $p(X_0 = x) = 1$ i.e. $p(X_0 \in S \setminus \{x\}) = 0$. Then we denote

$$P^t(x, A) = P(X_t \in A | X_0 = x)$$

for any $A \subset S$

Theorem Let $X_t, t \geq 0$ be a stationary, reversible Markov chain on the finite-state space S, with π as the stationary distribution. Let λ be the second largest (in modulus) eigenvalue of the transition matrix P. Then

(a) For all $t \geq 1$ and for any $i \in S$:

$$\sup_A \left| P^t(i, A) - \pi(A) \right| \leq \sqrt{\frac{1 - \pi_i}{\pi_i}} \frac{|\lambda|^t}{2}$$

Sampling methods

(b) For all $t \geq 1$ and any $i \in S$:

$$\sup_A \left| P^t(i, A) - \pi(A) \right| \leq \sqrt{\frac{p_{ii}^{(2)}}{\pi_i}} |\lambda|^{t-1}$$

where $p_{ii}^{(2)}$ is the ith diagonal entry of P^2

Sampling methods

(b) For all $t \geq 1$ and any $i \in S$:

$$\sup_{A} \left| P^t(i, A) - \pi(A) \right| \leq \sqrt{\frac{p_{ii}^{(2)}}{\pi_i}} |\lambda|^{t-1}$$

where $p_{ii}^{(2)}$ is the ith diagonal entry of P^2

(c) For all $t \geq 1$ and any initial distribution μ :

$$\chi^2(\mu^{(t)}, \pi) \leq |\lambda|^{2t} \chi^2(\mu, \pi)$$

Sampling methods

(b) For all $t \geq 1$ and any $i \in S$:

$$\sup_A \left| P^t(i, A) - \pi(A) \right| \leq \sqrt{\frac{p_{ii}^{(2)}}{\pi_i}} |\lambda|^{t-1}$$

where $p_{ii}^{(2)}$ is the ith diagonal entry of P^2

(c) For all $t \geq 1$ and any initial distribution μ :

$$\chi^2(\mu^{(t)}, \pi) \leq |\lambda|^{2t} \chi^2(\mu, \pi)$$

(d) For all $t \geq 1$ and any initial distribution μ :

$$\sup_A |P_\mu(X_t \in A) - \pi(A)| \leq \frac{|\lambda|^t}{2} \sqrt{\chi^2(\mu, \pi)}$$

Sampling methods

Question What is the second largest eigenvalue for the Metropolis–Hastings algorithm?

Sampling methods

Question What is the second largest eigenvalue for the Metropolis–Hastings algorithm?

Recall

$$p_{ij} = c \min\left\{1, \frac{\pi_j}{\pi_i}\right\}, j \neq i \text{ and } p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

Sampling methods

Question What is the second largest eigenvalue for the Metropolis–Hastings algorithm?

Recall

$$p_{ij} = c \min\left\{1, \frac{\pi_j}{\pi_i}\right\}, j \neq i \text{ and } p_{ii} = 1 - \sum_{j \neq i} p_{ij}$$

Eigenvalues of transition matrix corresponding to MH algo Suppose there are k states. Then setting $c = 1/k$, it can be shown that: if we label the states such that $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_k$ then $\lambda_1 = 1$ and

$$\lambda_l = \frac{1}{k}\left[\sum_{j=l-1}^{k} \frac{\pi_{l-1} - \pi_j}{\pi_{l-1}}\right], l \geq 2$$

Homework Verify the formula for small values of k.